Athens University of Economics and Business

School of Business

Department of Management Science & Technology

Master of Science in Business Analytics

Program:	Full-time
Quarter:	1st (Fall Quarter)
Course:	Statistics for Business Analytics I
Assignment №:	2
Students (Registration №):	Souflas Eleftherios-Efthymios (f2822217)

Statistics for Business Analytics I
Lab Assignment #2

Student: Souflas Eleftherios-Efthymios

Question 1

Read the "usdata" dataset and use str() to understand its structure.

Output

```
# install.packages("foreign")
file <- "path\\to\\usdata"
library(foreign)
usdata <- read.csv(file, header = TRUE, sep = " ", quote = "\"")
str(usdata)
```

Comment

Usdata is a data frame containing 63 observations (cases) from the files of a big real estate agency in USA, concerning house sales from 15th of February till 30th of April 1993. Each observation consists of the following 6 variables:

- PRICE, an integer vector, containing the price in hundreds of dollars the house was sold.
- SQFT, an integer vector, containing the house's living space in square feet.
- AGE, an integer vector, containing the house's age in years (how many years before 1993 was constructed).
- FEATS, an integer vector, containing how many features in quantity (e.g., 1, 4, 6, etc.) the house has from a pool of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access).
- NE, an integer vector, mentioning if the house is located in the northeast sector of the city (value = 1) or not (value = 0).
- COR, an integer vector, mentioning if the house is placed at a corner location (value = 1) or not (value = 0).

Question 2

Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and NE, COR to be factors.

Output

```
for (i in 1:4){
  usdata[,i] <- as.numeric(as.character(usdata[,i]))
}
for (i in 5:6){
  usdata[,i] <- as.factor(usdata[,i])
}
str(usdata)
```

```
> str(usdata)
'data.frame':    63 obs. of  6 variables:
 $ PRICE: num  2050 2150 2150 1999 1900 ...
 $ SQFT : num  2650 2664 2921 2580 2580 ...
 $ AGE  : num  3 28 17 20 20 10 2 2 20 30 ...
 $ FEATS: num  7 5 6 4 4 4 5 3 5 6 ...
 $ NE   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ COR  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Comment

Two “for” loops are used in order to iterate over the variables of the data frame and convert them accordingly to the proper data type.

Question 3

Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings.

Output

```
# convert hundreds of dollars to dollars
usdata$PRICE <- usdata$PRICE*100
# convert square feet of living space to square meters
usdata$SQFT <- usdata$SQFT/10.764
colnames(usdata)[2] <- "SQMT"
summary(usdata)
```

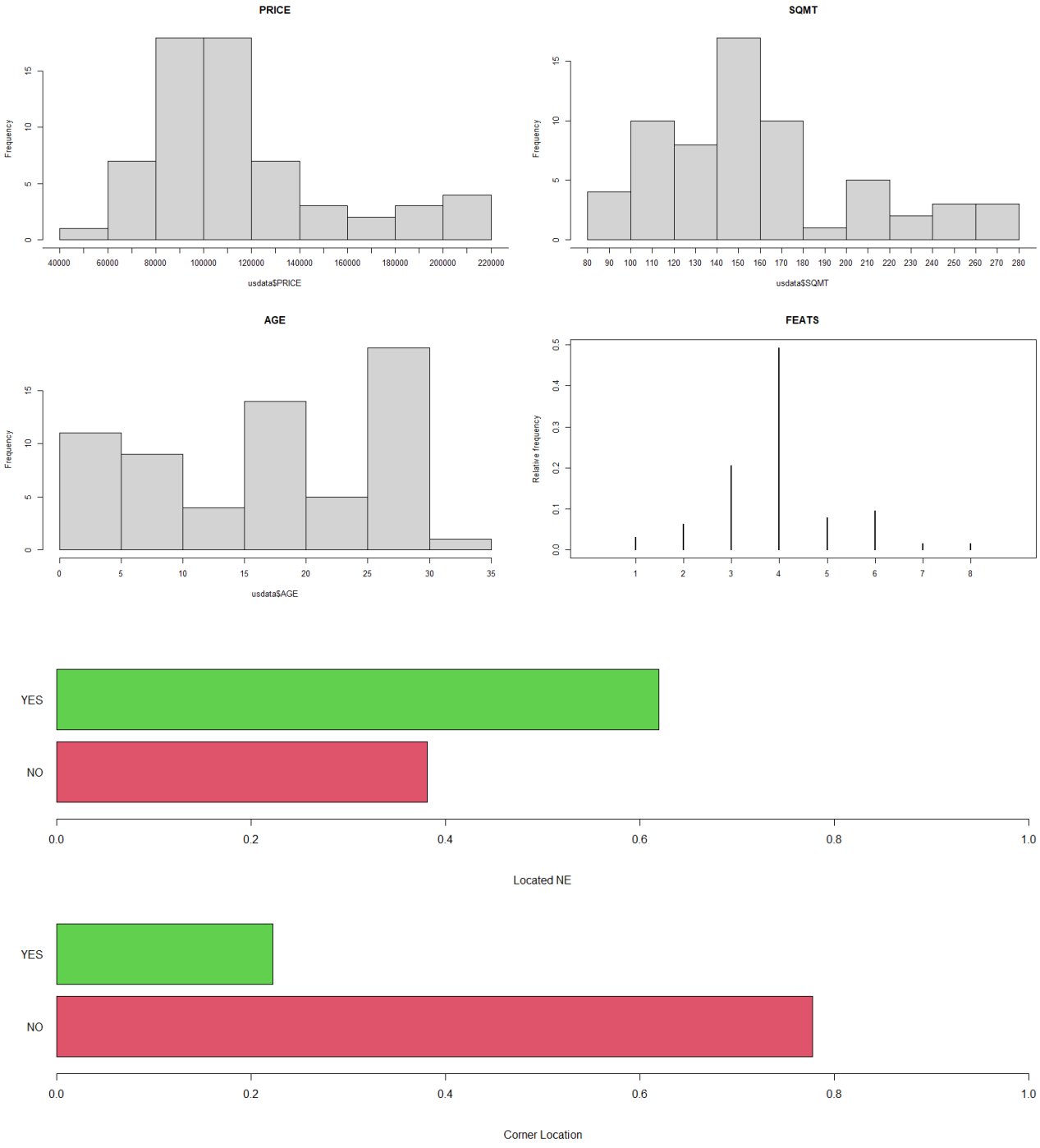
```
> summary(usdata)
```

	PRICE	SQMT	AGE	FEATS	NE	COR
Min.	58000	90.12	2.00	1.000	0: 24	0: 49
1st Qu.	91000	130.06	7.00	3.000	1: 39	1: 14
Median	104900	156.08	20.00	4.000		
Mean	115841	160.68	17.46	3.952		
3rd Qu.	125000	178.37	27.50	4.000		
Max.	215000	272.30	31.00	8.000		

```
library(psych)
round(t(describe(usdata[,1:4])),2)
```

```
> round(t(describe(usdata[,1:4])),2)
```

	PRICE	SQMT	AGE	FEATS
Vars	1.00	2.00	3.00	4.00
N	63.00	63.00	63.00	63.00
Mean	115841.27	160.68	17.46	3.95
Sd	39270.88	47.07	9.60	1.28
Median	104900.00	156.08	20.00	4.00
Trimmed	110596.08	156.56	17.75	3.92
Mad	26242.02	36.50	11.86	1.48
Min	58000.00	90.12	2.00	1.00
Max	215000.00	272.30	31.00	8.00
Range	157000.00	182.18	29.00	7.00
Skew	1.18	0.74	-0.21	0.45
Kurtosis	0.54	-0.16	-1.47	1.12
Se	4947.67	5.93	1.21	0.16



Comment

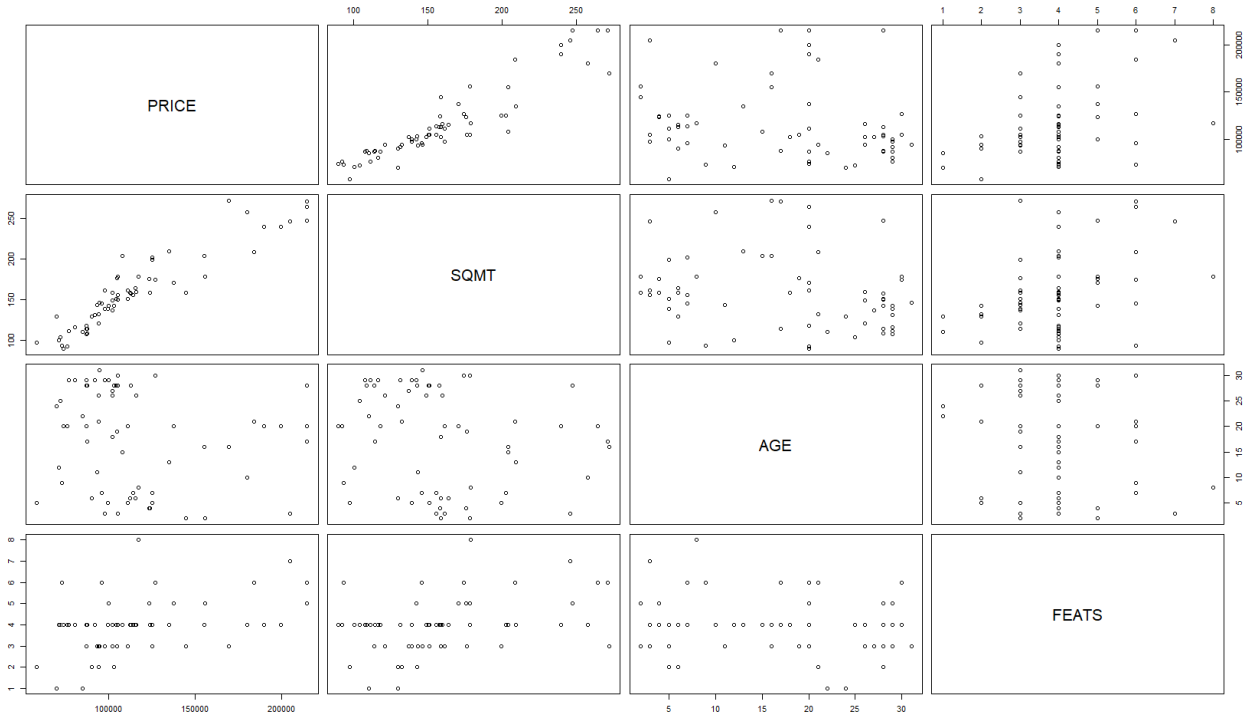
We converted the price that the houses were sold from hundreds of dollars to dollars in order to have a better understanding of the variable and to avoid misleading assumptions to be made. The same we did with the house's living space. We transformed it from square feet to square meters as it is the metric unit that we can understand better, changing the name of the variable at the same time to avoid creating confusion. We can observe in our data that there not exist null or missing values for any of the variables. The price the houses were sold range from 58 thousand dollars to 215 thousand dollars, whereas most of the houses were sold for a range of approximately 80-120 thousand dollars. The living space that the houses sold occupy range from 90 to 272 square meters and most often houses of approximately 140 to 160 square meters were sold. The age of houses sold vary from 2 to 31 years with most houses sold falling in the range of 25-30 years. That means that they were houses constructed in 1963 till 1968. All houses sold had at least 1 and at most 8 out of the 11 features and most of them had 4 out of 11 features. Finally, 60% of the houses sold were located at the Northeast sector of the city and 20% of the whole 63 cases studied were placed at a corner

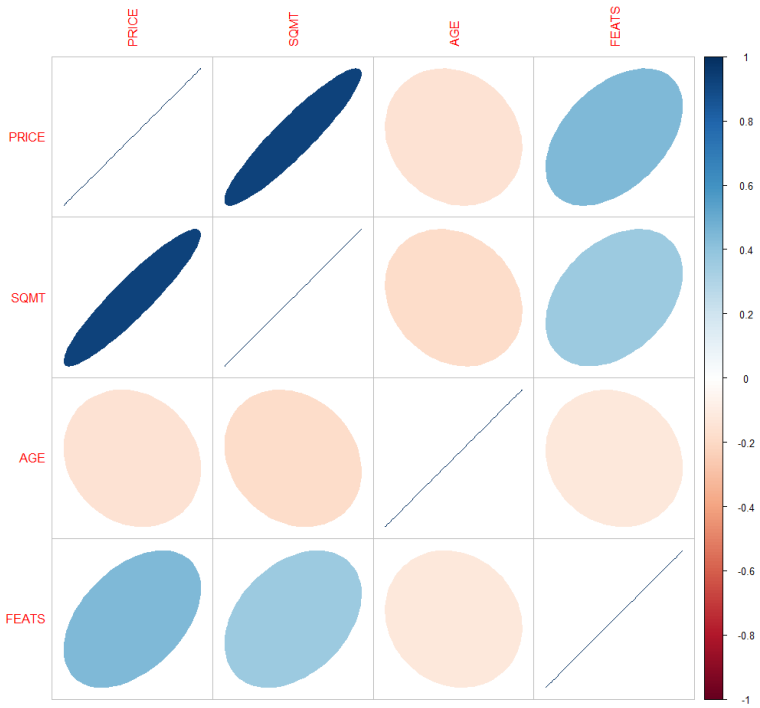
location.

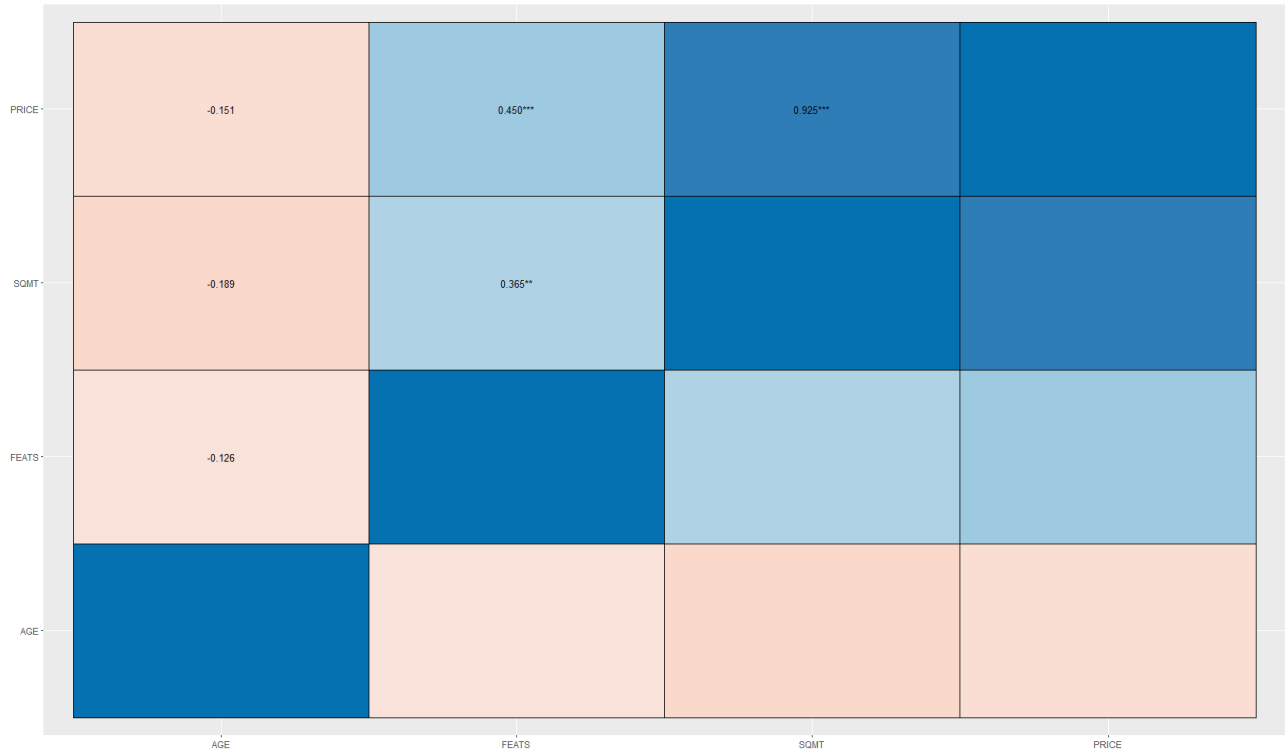
Question 4

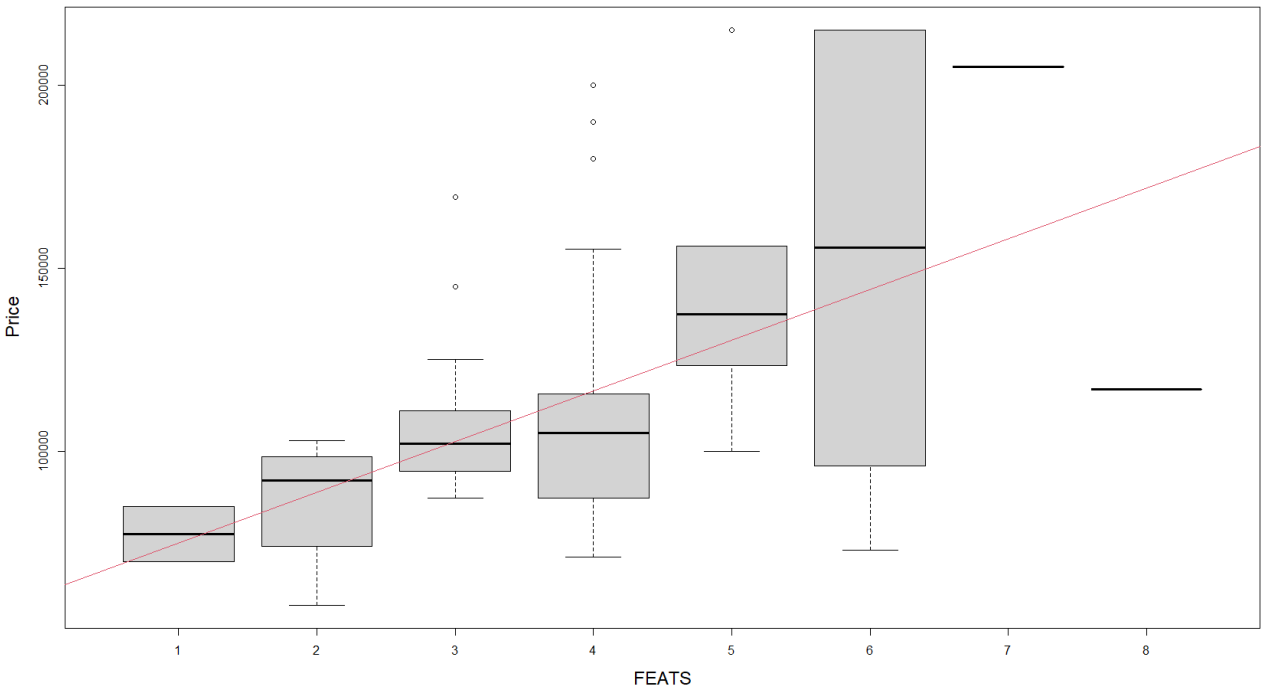
Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset. (Hint: Plot variables against one another and use correlation plots and measures for the numerical variables). Comment on your findings. Is there a linear relationship between PRICE and any of the variables in the dataset?

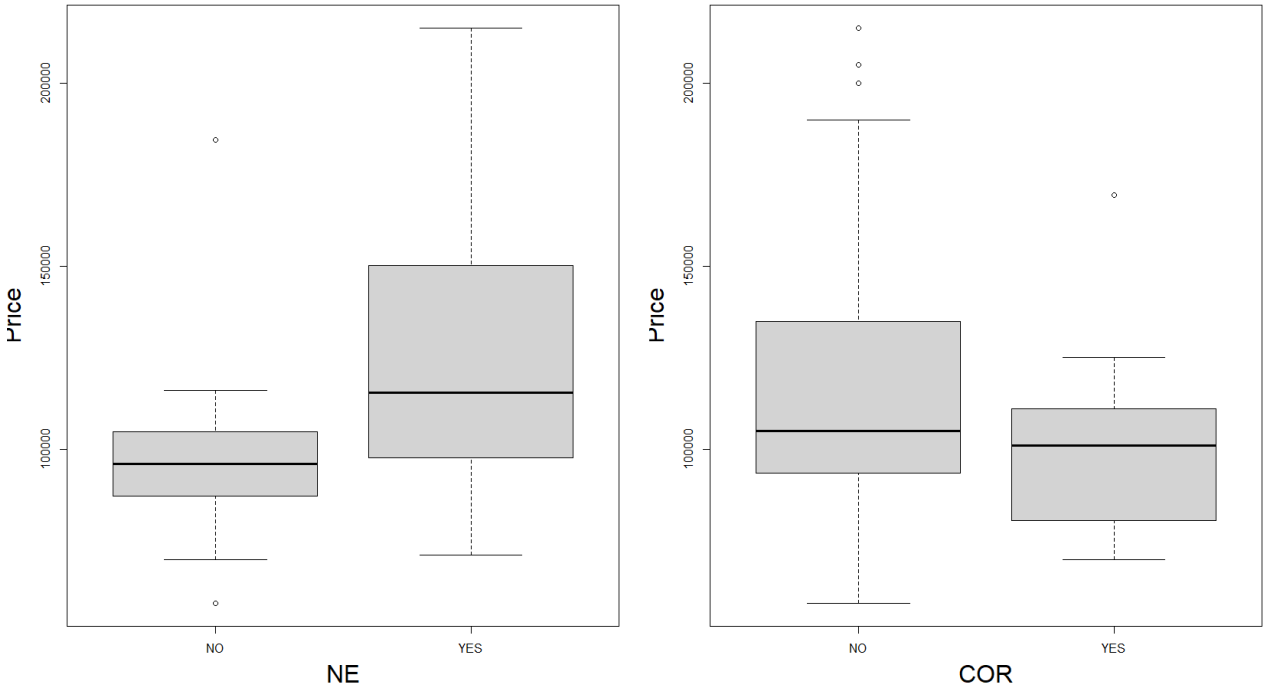
Output






PRICE
SQMT
FEATS
AGE
-0.151
0.450***
0.925***
-0.189
0.365**
-0.126
AGE
FEATS
SQMT
PRICE


Price
100000
150000
200000
1
2
3
4
5
6
7
8
FEATS




```
round(cor(usdata[,1:4]),2)
```

	PRICE	SQMT	AGE	FEATS
PRICE	1.00	0.93	-0.15	0.45
SQMT	0.93	1.00	-0.19	0.36
AGE	-0.15	-0.19	1.00	-0.13
FEATS	0.45	0.36	-0.13	1.00

```
cor(usdata$PRICE, usdata$SQMT, method = "pearson")
```

[1] 0.9251753

Comment

From the conduction of the pairwise comparisons between the variables in the dataset, a strong association between PRICE and SQMT is implied by the dataset. From the scatterplot between each pair of numeric variables in the dataset, we can imply a strong positive linear relationship, as it was expected, between the price of a house and its living space. So, as the living space of a house tends to get larger, the price the house was sold gets larger too. The Pearson correlation coefficient also comes to strengthen our previous implication as it measures an R = 0.93 between the aforementioned variables. There also seems to be a bivariate association between price and feats of a house. The more features a house has, the more it seems to cost. But a strong linear relationship cannot be implied as the median price of the houses do not seem to fall over a line as the features a house has increase. The Pearson correlation coefficient measures an R = 0.45 (medium linear dependence) between PRICE and FEATS variables. Other possible bivariate linear relationships between any other variables cannot be assessed.

Question 5

Construct a model for the expected selling prices (PRICE) according to the remaining features. (Hint: Conduct multiple regression having PRICE as a response and all the other variables as predictors). Does this linear model fit well to the data? (Hint: Comment on R^2 adj.).

Output

```
model <- lm(PRICE ~., data = usdata)
summary(model)
# R^2 is above 0.7, so current linear model is a good fit model, but not
a very good (below 0.9)
```

```
> summary(model)

Call:
lm(formula = PRICE ~ ., data = usdata)

Residuals:
   Min     1Q Median     3Q    Max
-41611  -7103  -1526   8302  34777

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19334.93    9452.38  -2.046   0.0454 *
SQMT            728.31      44.11  16.509   <2e-16 ***
AGE             222.91     228.63   0.975   0.3337
FEATS          3436.57    1627.11   2.112   0.0391 *
NEYES          3000.45    4793.94   0.626   0.5339
CORYES        -5307.94    4615.65  -1.150   0.2550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14480 on 57 degrees of freedom
Multiple R-squared:  0.8749,     Adjusted R-squared:  0.864
F-statistic: 79.76 on 5 and 57 DF,  p-value: < 2.2e-16
```

Comment

We constructed a multiple regression model for the expected selling prices of houses (response) according to all other variables (predictors). We can see that variables AGE, NEYES and CORYES are not significant predictors of our response variable (PRICE). The Adjusted R-Squared is 0.864. That means that 86.4% of variance is explained by our model. The model we constructed for the prediction of the price of a house is a good fit with regards to the data we have.

Question 6

Find the best model for predicting the selling prices (PRICE). Select the appropriate features using stepwise methods. (Hint: Use Forward, Backward or Stepwise procedure according to AIC or BIC to choose which variables appear to be more significant for predicting selling PRICES).

Output

```
mfull <- lm(PRICE~.,data=usdata)
step(mfull, direction='both')
```

```
> step(mfull, direction='both')
Start:  AIC=1212.87
PRICE ~ SQMT + AGE + FEATS + NE + COR

       Df  Sum of Sq       RSS    AIC
- NE    1 8.2178e+07 1.2040e+10 1211.3
- AGE   1 1.9942e+08 1.2157e+10 1211.9
- COR   1 2.7743e+08 1.2235e+10 1212.3
<none>              1.1958e+10 1212.9
- FEATS 1 9.3580e+08 1.2893e+10 1215.6
- SQMT  1 5.7178e+10 6.9136e+10 1321.4

Step:  AIC=1211.31
PRICE ~ SQMT + AGE + FEATS + COR
```

```
      Df  Sum of Sq       RSS    AIC
- AGE    1 1.2171e+08 1.2161e+10 1209.9
- COR    1 2.5099e+08 1.2291e+10 1210.6
<none>               1.2040e+10 1211.3
+ NE     1 8.2178e+07 1.1958e+10 1212.9
- FEATS  1 1.0695e+09 1.3109e+10 1214.7
- SQMT   1 6.2889e+10 7.4928e+10 1324.5

Step:  AIC=1209.94
PRICE ~ SQMT + FEATS + COR

      Df  Sum of Sq       RSS    AIC
- COR    1 2.2454e+08 1.2386e+10 1209.1
<none>               1.2161e+10 1209.9
+ AGE    1 1.2171e+08 1.2040e+10 1211.3
+ NE     1 4.4654e+06 1.2157e+10 1211.9
- FEATS  1 1.0426e+09 1.3204e+10 1213.1
- SQMT   1 6.3520e+10 7.5682e+10 1323.1

Step:  AIC=1209.09
PRICE ~ SQMT + FEATS

      Df  Sum of Sq       RSS    AIC
<none>               1.2386e+10 1209.1
+ COR    1 2.2454e+08 1.2161e+10 1209.9
+ AGE    1 9.5255e+07 1.2291e+10 1210.6
+ NE     1 2.1763e+06 1.2384e+10 1211.1
- FEATS  1 1.3876e+09 1.3774e+10 1213.8
- SQMT   1 6.3899e+10 7.6285e+10 1321.6

Call:
lm(formula = PRICE ~ SQMT + FEATS, data = usdata)

Coefficients:
(Intercept)        SQMT       FEATS
  -17592.8       732.5      3983.7
```

```
step(mfull, direction='both', k=log(100))
> step(mfull, direction='both', k=log(100))
Start:  AIC=1228.51
PRICE ~ SQMT + AGE + FEATS + NE + COR

      Df  Sum of Sq       RSS    AIC
- NE     1 8.2178e+07 1.2040e+10 1224.3
- AGE    1 1.9942e+08 1.2157e+10 1224.9
- COR    1 2.7743e+08 1.2235e+10 1225.3
<none>               1.1958e+10 1228.5
- FEATS  1 9.3580e+08 1.2893e+10 1228.7
- SQMT   1 5.7178e+10 6.9136e+10 1334.5

Step:  AIC=1224.33
PRICE ~ SQMT + AGE + FEATS + COR
```

```
      Df  Sum of Sq       RSS    AIC
- AGE    1 1.2171e+08 1.2161e+10 1220.4
- COR    1 2.5099e+08 1.2291e+10 1221.0
<none>               1.2040e+10 1224.3
- FEATS  1 1.0695e+09 1.3109e+10 1225.1
+ NE     1 8.2178e+07 1.1958e+10 1228.5
- SQMT   1 6.2889e+10 7.4928e+10 1334.9

Step:  AIC=1220.36
PRICE ~ SQMT + FEATS + COR

      Df  Sum of Sq       RSS    AIC
- COR    1 2.2454e+08 1.2386e+10 1216.9
<none>               1.2161e+10 1220.4
- FEATS  1 1.0426e+09 1.3204e+10 1220.9
+ AGE    1 1.2171e+08 1.2040e+10 1224.3
+ NE     1 4.4654e+06 1.2157e+10 1224.9
- SQMT   1 6.3520e+10 7.5682e+10 1330.9

Step:  AIC=1216.91
PRICE ~ SQMT + FEATS

      Df  Sum of Sq       RSS    AIC
<none>               1.2386e+10 1216.9
- FEATS  1 1.3876e+09 1.3774e+10 1219.0
+ COR    1 2.2454e+08 1.2161e+10 1220.4
+ AGE    1 9.5255e+07 1.2291e+10 1221.0
+ NE     1 2.1763e+06 1.2384e+10 1221.5
- SQMT   1 6.3899e+10 7.6285e+10 1326.8

Call:
lm(formula = PRICE ~ SQMT + FEATS, data = usdata)

Coefficients:
(Intercept)        SQMT       FEATS
  -17592.8       732.5      3983.7
```

Comment

We can see that models selected with stepwise method from the full model both from AIC (default) and BIC (k = log(100)) agree on the predictors (SQMT and FEATS) selection and on the coefficients' value. The variables that appear to be more significant for predicting selling PRICES (as assumed earlier) are the living space of the house (SQMT) and the number of features it contains (FEATS).

Question 7

Get the summary of your final model, (the model that you ended up having after conducting the stepwise procedure) and comment on the output. Interpret the coefficients. Comment on the significance of each coefficient and write down the mathematical formulation of the model (e.g PRICES = Intercept + coef1*Variable1 + coef2*Variable2 +…. + ε where ε ~ N(0, ...)). Should the intercept be excluded from our model?

Output

```
model2 <- lm(PRICE ~ 1 + SQMT + FEATS, data = usdata)
```

```
# model2 <- lm(PRICE ~ . - AGE - NE - COR, data = usdata)
summary(model2)
```

> summary(model2)

Call:
lm(formula = PRICE ~ 1 + SQMT + FEATS, data = usdata)

Residuals:
  Min 1Q Median 3Q Max
-40044 -7170 -1121 9312 34182

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) -17592.76 7434.21 -2.366 0.0212 *
SQMT 732.45 41.63 17.594 <2e-16 ***
FEATS 3983.69 1536.53 2.593 0.0119 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14370 on 60 degrees of freedom
Multiple R-squared: 0.8705, Adjusted R-squared: 0.8661
F-statistic: 201.6 on 2 and 60 DF, p-value: < 2.2e-16

```
round(sapply(usdata[,c(2,4)], mean),2)
```

 SQMT FEATS
160.68 3.95

```
round(sapply(usdata[,c(2,4)], sd),2)
```

 SQMT FEATS
47.07 1.28

```
47.07  1.28

usdata2 <- as.data.frame(scale(usdata[,1:4], center = TRUE, scale = F))
usdata2$NE <- usdata$NE
usdata2$COR <- usdata$COR
usdata2$PRICE <- usdata$PRICE
round(sapply(usdata2[,c(2,4)], mean),2)
round(sapply(usdata2[,c(2,4)], sd),2)
mfull2 <- lm(PRICE~.,data=usdata2)
step(mfull2, direction='both')
model3 <- lm(PRICE ~ 1 + SQMT + FEATS, data = usdata2)
summary(model3)
```

> summary(model3)

Call:
lm(formula = PRICE ~ 1 + SQMT + FEATS, data = usdata2)

Residuals:
  Min 1Q Median 3Q Max
-40044 -7170 -1121 9312 34182

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 115841.27    1810.17  63.995   <2e-16 ***
SQMT          732.45      41.63  17.594   <2e-16 ***
FEATS        3983.69    1536.53   2.593   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14370 on 60 degrees of freedom
Multiple R-squared:  0.8705,     Adjusted R-squared:  0.8661
F-statistic: 201.6 on 2 and 60 DF,  p-value: < 2.2e-16
```

Comment

We got the summary of the final model, after the conduction of the stepwise method. We get an Adjusted R-Squared of 0.8661. That means that 86.61% of variance is explained by our model, which is slightly better than our previous model. The mathematical formulation of the model is:

$$(PRICES) = -17592.76 + 732.45 * (SQMT) + 3983.69 * (FEATS) + \varepsilon, \text{ where } \varepsilon \sim N(0, 14370^2)$$

If we try to interpret the coefficients, we observe that when the house has zero size (SQMT) and no features, then the expected value of it is equal to -17592.76 dollars. This is considered as fixed cost (which is frequent in Economics) but the interpretation is not sensible. Because it is significant it should not be excluded from our model. Instead, we should try to fit the model with centred to zero covariates. Thus, we produce a third model (model3) and when we get the summary of it we observe the following:

A. The coefficients of SQMT and FEATS have not changed, as we just centred our point of reference, but the data are the same.
B. σ^2 has the same value for the same reason as before.
C. R_{adj}^2 has the same value.
D. The (Intercept) β_0 value has changed and its value has become more significant for our model.

The mathematical formulation of the new model is:

$$(PRICES) = 115841.27 + 732.45 * (SQMT) + 3983.69 * (FEATS) + \varepsilon, \text{ where } \varepsilon \sim N(0, 14370^2)$$

Interpreting the coefficients, we observe that:

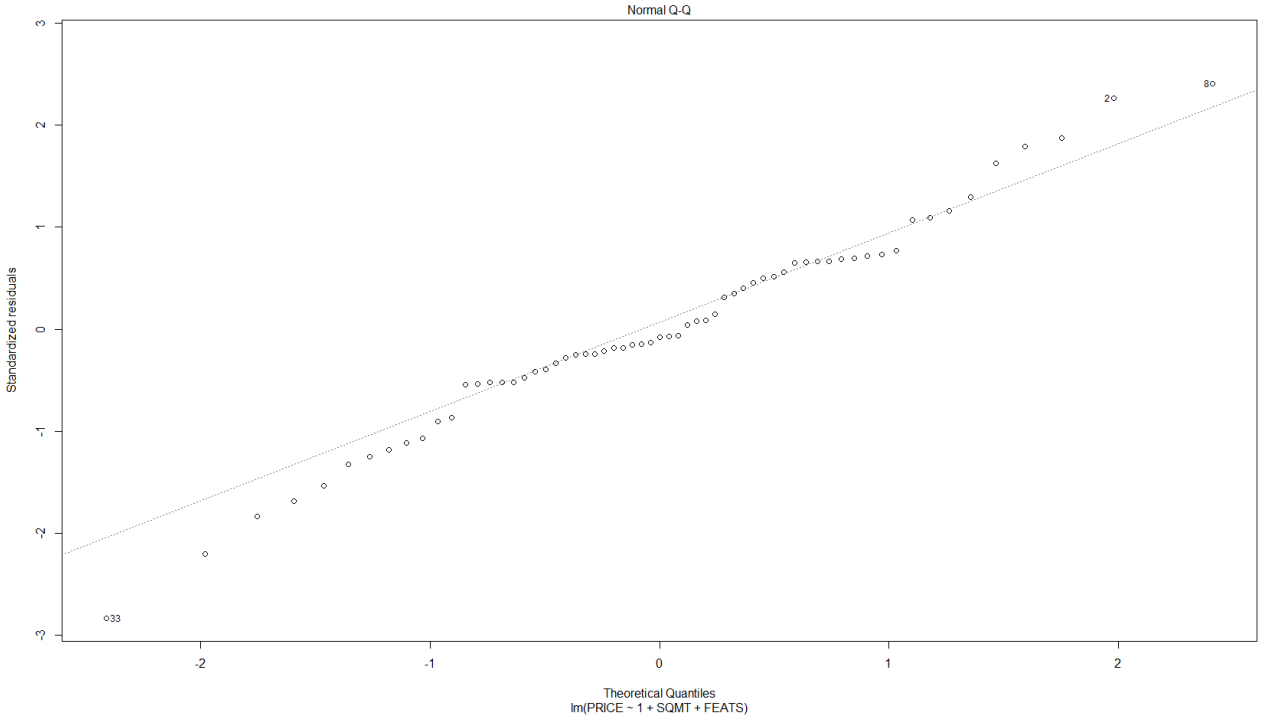
A. We need approximately 115841 dollars to buy a house of average size (161 m^2) that has 4 out of 11 features.
B. If we compare two houses with the same characteristics which differ only by 1 m^2, then the expected difference in the price will be 732.45 $ in favour of the larger house.
C. If we compare two houses with the same characteristics which differ only by 1 feature, then the expected difference in the price will be 3983.69 $ in favour of the house with more features.

Question 8

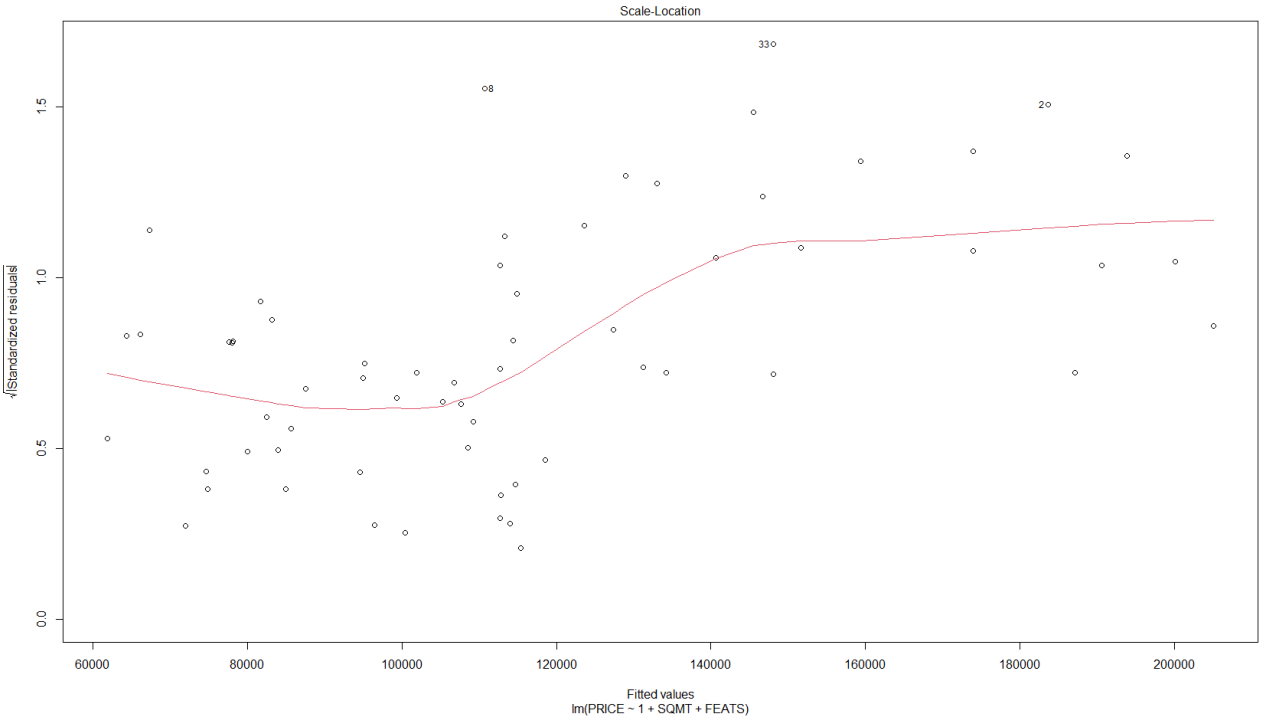
Check the assumptions of your final model. Are the assumptions satisfied? If not, what is the impact of the violation of the assumption not satisfied in terms of inference? What could someone do about it?

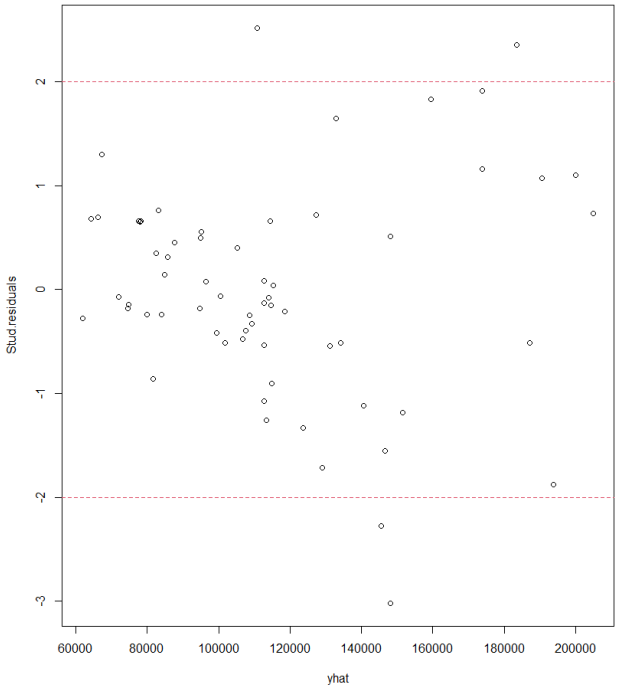
Output

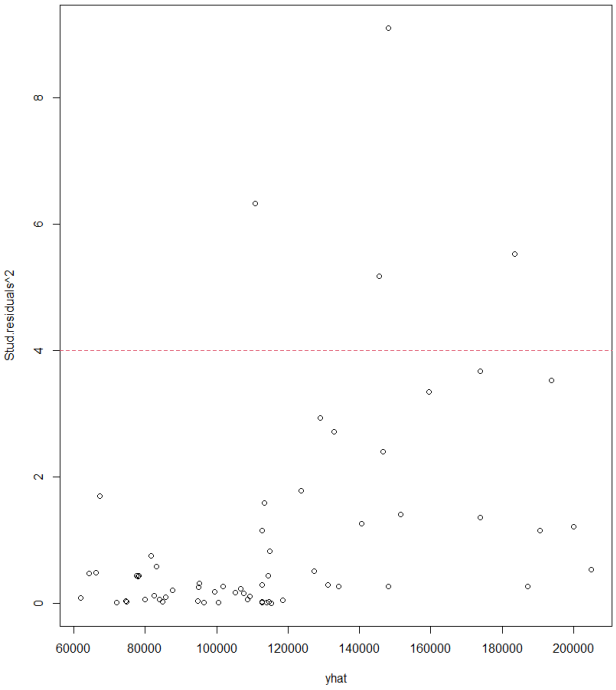
```
# Normality of errors (Q-Q Plot)
```



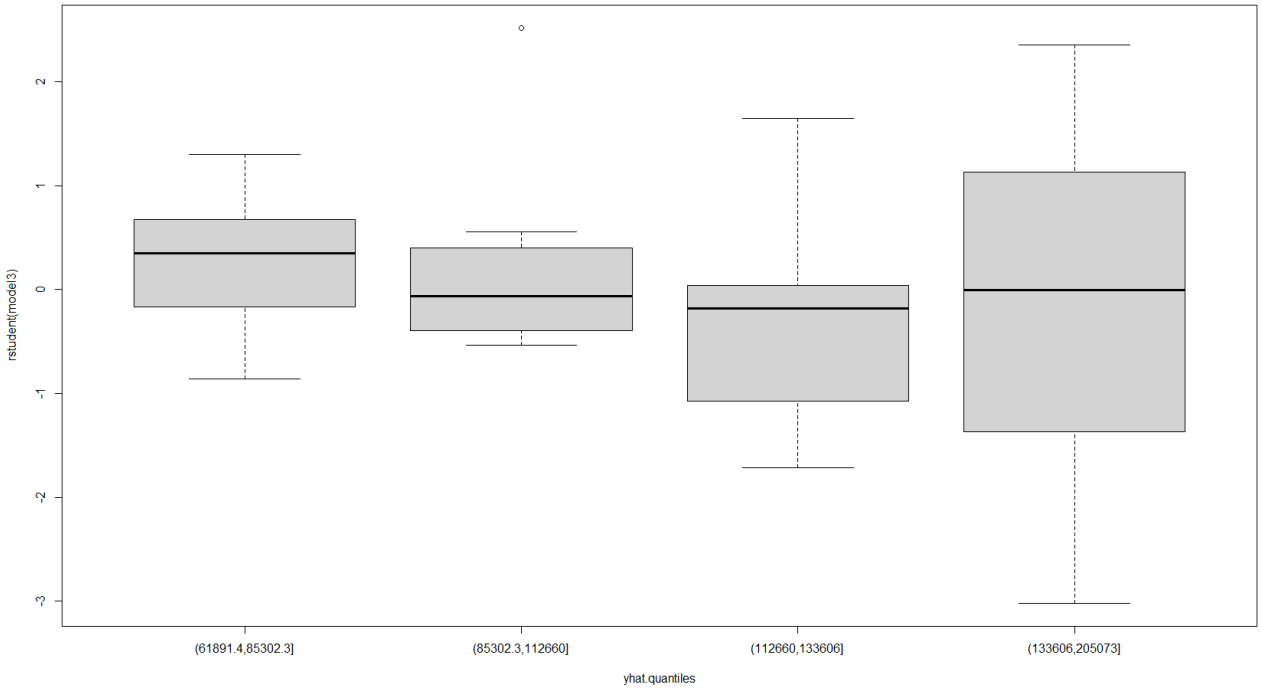

```
# Constant Variance (Homoscedasticity)
```



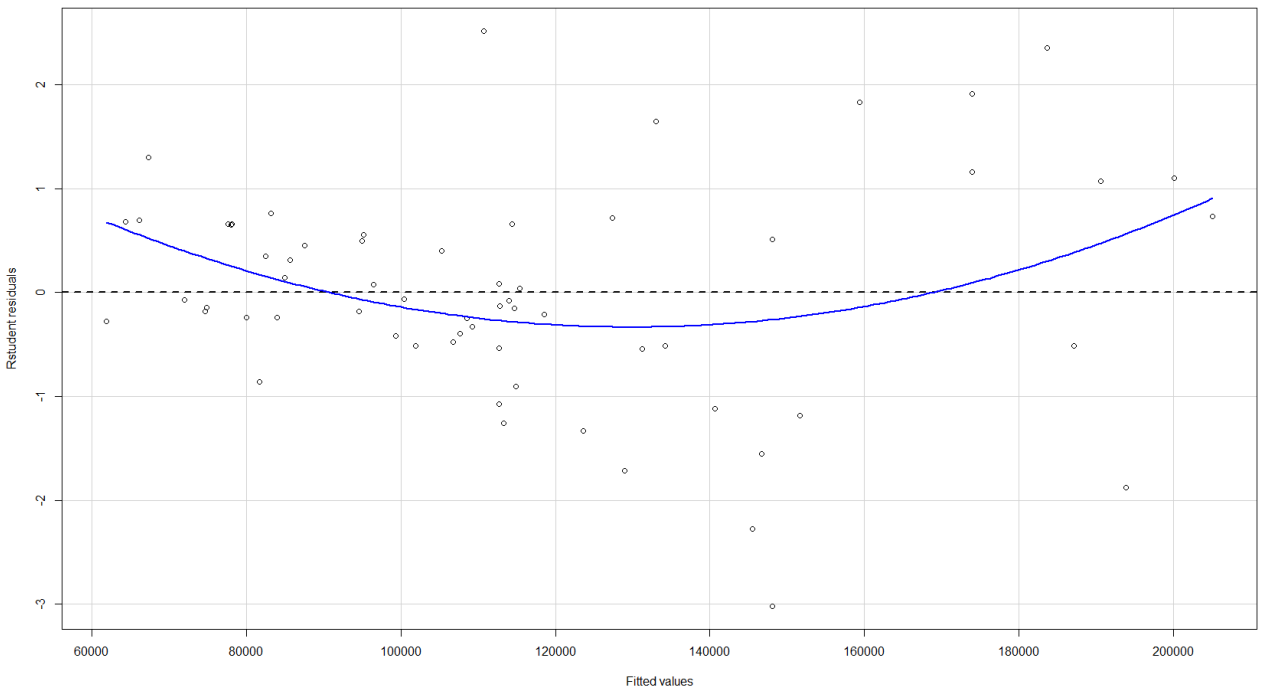




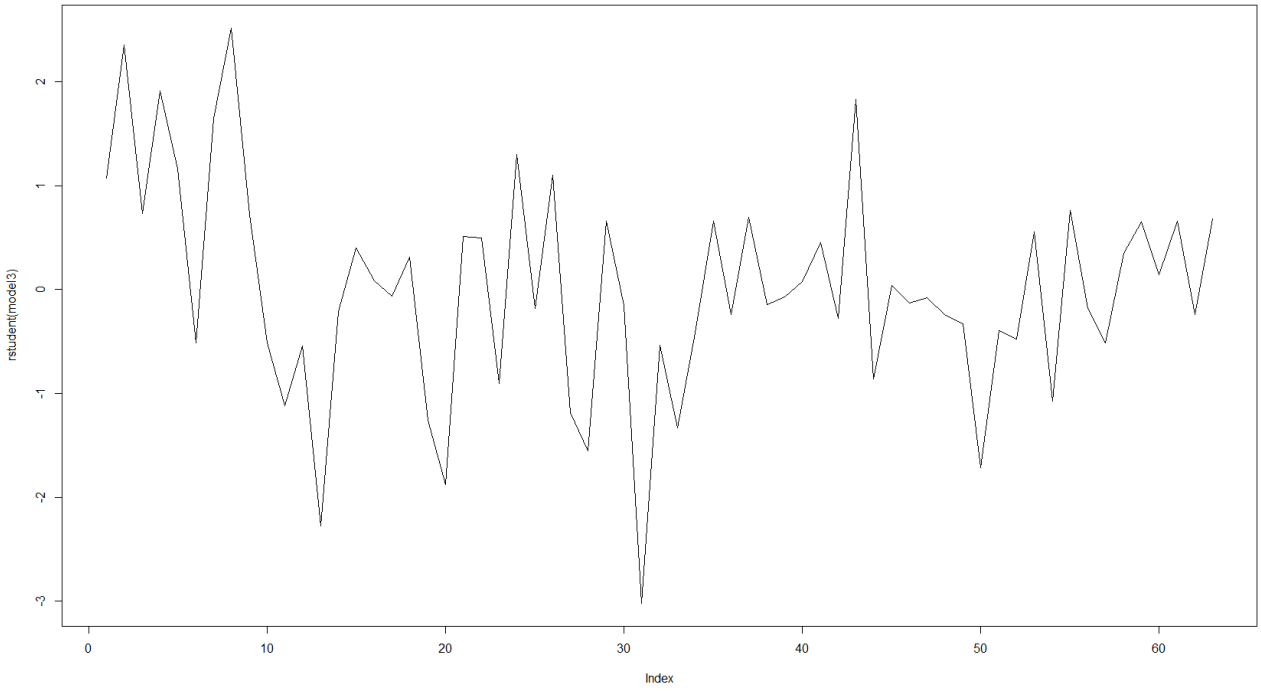

```
> ncvTest(model3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 14.99402, Df = 1, p = 0.00010785
> yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab=6)
> table(yhat.quantiles)
yhat.quantiles
(61891.4,85302.3]  (85302.3,112660]   (112660,133606]   (133606,205073]
            15              17              14              16
> leveneTest(rstudent(model3)~yhat.quantiles)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  3  9.9191 2.249e-05 ***
      58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




```
# Linearity of response and predictors
> residualPlots(model3, plot=F, type = "rstudent")
          Test stat Pr(>|Test stat|)
SQMT         2.0388        0.045959 *
FEATS       -0.2876        0.774643
Tukey test   2.6002        0.009317 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
# Independence of Errors
```




```
> library(randtests)
> randtests::runs.test(model3$res)

        Runs Test

data:  model3$res
statistic = -0.25611, runs = 31, n1 = 31, n2 = 31, n = 62, p-value = 0.7979
alternative hypothesis: nonrandomness

> library(lmtest)
> dwtest(model3)

        Durbin-Watson test

data:  model3
DW = 1.5734, p-value = 0.03571
alternative hypothesis: true autocorrelation is greater than 0

> library(car)
> durbinWatsonTest(model3)
 lag Autocorrelation D-W Statistic p-value
   1      0.2012826      1.573363   0.086
 Alternative hypothesis: rho != 0
```

Comment

None of the assumptions is satisfied, except the independence of errors, because the assumption of randomness is not rejected (Runs Test), the autocorrelation of the disturbances to be 0 could not be rejected (Durbin-Watson Test) and mainly because the data have no meaning in terms of time sequence. The analysis of independence of errors should be skipped since it is not possible to check for independence.

The violation of the normality assumption leads the performance of hypothesis tests and confidence

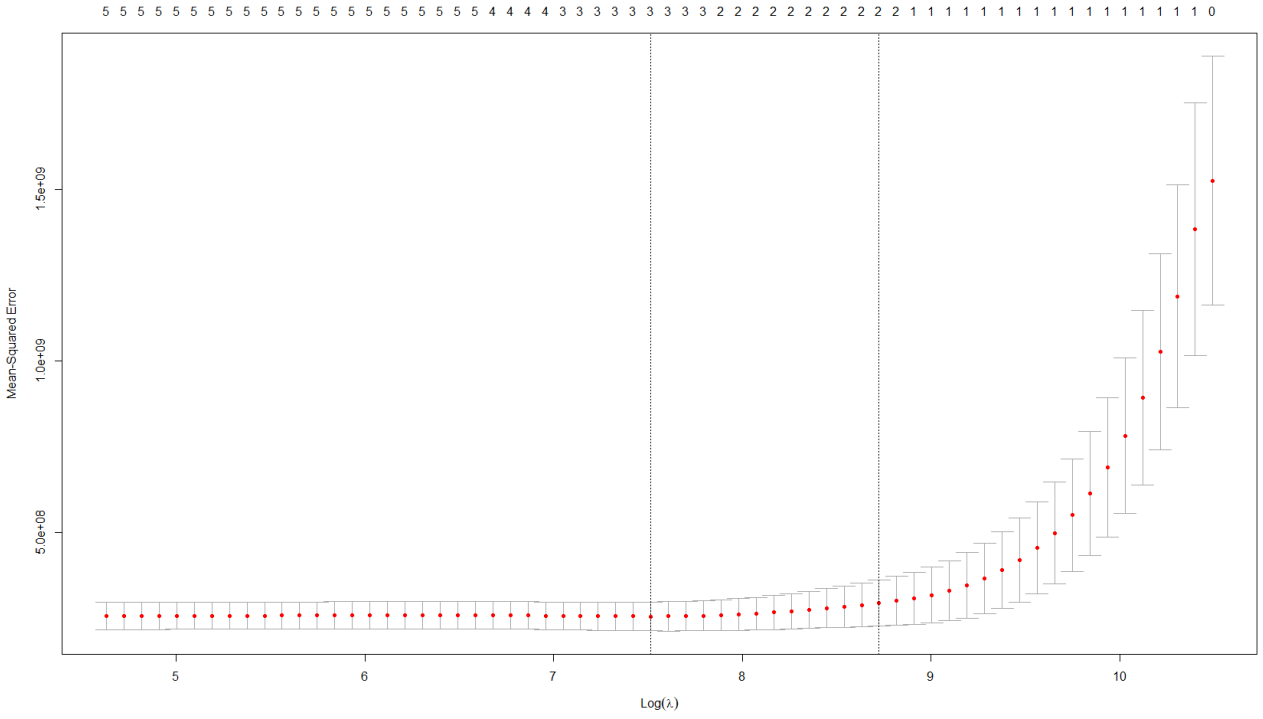
intervals to be compromised, although the aforementioned procedures are generally robust to small departures from Normality. In order to face the problem, we could use log or box-cox transformations, non-normal errors, GLM models for non-normal responses or non-parametric regression models.
The violation of the assumption of homoscedasticity of errors leads the estimators of coefficients to be unbiased, the error variance estimator not to be estimated correctly and the standard errors not to be estimated appropriately, affecting this way the performance of hypothesis tests and confidence intervals. To face this challenge, we could use weighted least squares regression models, transformed response, GLMs with more complicated distributions, GAMLSS to use covariates in the variance components.
The violation of linearity leads the error variance to appear as non-constant, even if it is constant, due to the model misspecification and the model to be inadequate, especially for prediction. To deal with this situation, we could transform the response, the covariates, use polynomial or non-parametric regression models or use non-linear models.

Question 9

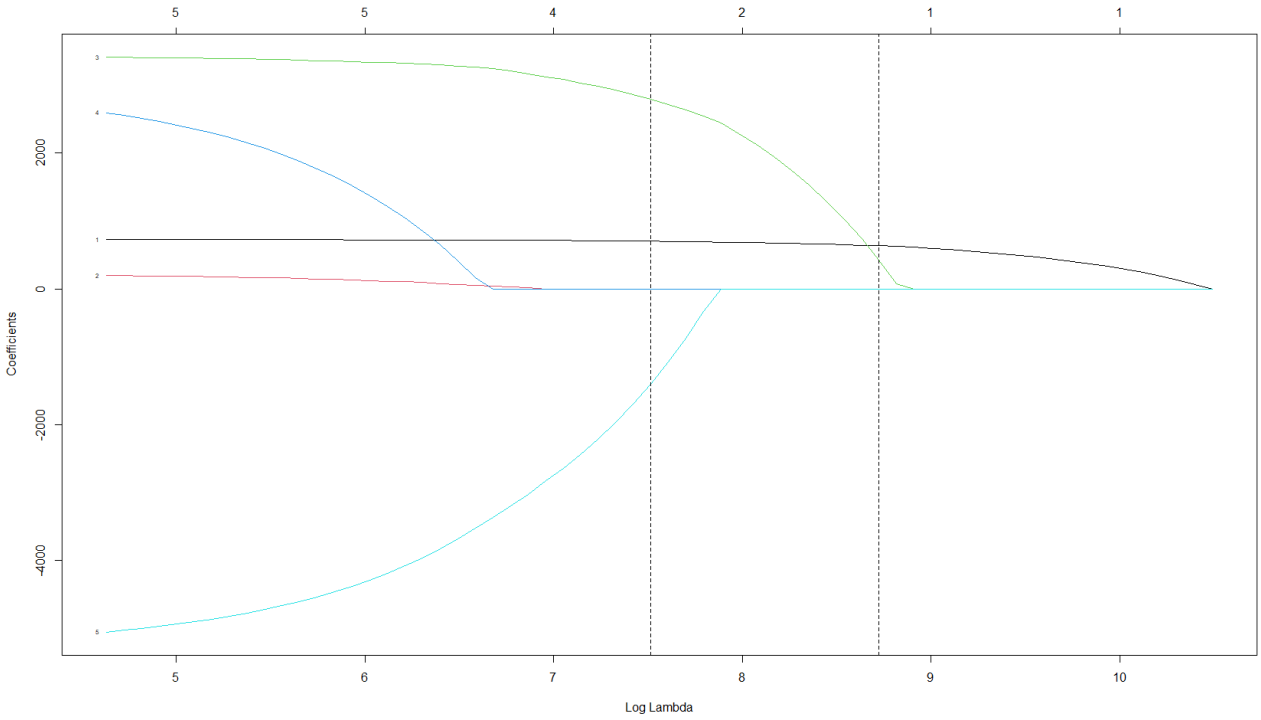
Conduct LASSO as a variable selection technique and compare the variables that you end up having using LASSO to the variables that you ended up having using stepwise methods in (VI). Are you getting the same results? Comment.

Output

```
> lasso1$lambda.min
[1] 1836.077
> lasso1$lambda.1se
[1] 6153.801
```




```
> coef(lasso1, s = "lambda.1se")
6 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 115841.2698
SQMT           635.8961
AGE              .
FEATS          419.9710
NEYES            .
CORYES           .
```



Comment

We conducted LASSO as a variable screening technique, and we found out that the variables we ended up having using LASSO were the same to those we had using the stepwise methods in Question 6 (Intercept, SQMT and FEATS). However, LASSO suggested another model:

$$(PRICES) = 115841.27 + 635.9 * (SQMT) + 419.97 * (FEATS) + \boldsymbol{\varepsilon}$$

β_0 coefficient has exactly the same value, whereas β_1 and β_2 coefficients have significantly smaller values. LASSO can become handy, not in this small case, but when p is large (p >> n), because it will clear all irrelevant variables very fast.

Code

```
#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 1                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# install.packages("foreign")
file <- "path\\to\\usdata"
library(foreign)
usdata <- read.csv(file, header = TRUE, sep = " ", quote = "\"")
str(usdata)

#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 2                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

for (i in 1:4){
  usdata[,i] <- as.numeric(as.character(usdata[,i]))
}
for (i in 5:6){
  usdata[,i] <- as.factor(usdata[,i])
}
str(usdata)


#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 3                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# convert hundreds of dollars to dollars
usdata$PRICE <- usdata$PRICE*100
# convert square feet of living space to square meters
usdata$SQFT <- usdata$SQFT/10.764
colnames(usdata)[2] <- "SQMT"
summary(usdata)
library(psych)
round(t(describe(usdata[,1:4])),2)
n <- nrow(usdata)
par(mfrow = c(2,2))
hist(usdata$PRICE, main = names(usdata)[1], xaxt="n")
axis(1, at = seq(0,max(usdata$PRICE)+sd(usdata$PRICE),by=10000))
hist(usdata$SQMT, main = names(usdata)[2], xaxt="n")
axis(1, at = seq(0,max(usdata$SQMT)+sd(usdata$SQMT),by=10))
hist(usdata$AGE, main = names(usdata)[3])
plot(table(usdata[,4])/n, type='h', xlim=range(usdata[,4])+c(-1,1), main=names(usdata)[4], ylab='Relative frequency')
levels(usdata$NE) = c("NO", "YES")
levels(usdata$COR) = c("NO", "YES")
par(mfrow = c(2,1))
barplot(table(usdata$NE)/n, horiz=T, las=1, col=2:3, xlim=c(0,1), ylim=c(0,2), sub = "Located NE",
        names.arg = levels(usdata$NE))
barplot(table(usdata$COR)/n, horiz=T, las=1, col=2:3, xlim=c(0,1), ylim=c(0,2), sub = "Corner Location",
        names.arg = levels(usdata$COR))


#########################################################################################################################################
#########################################################################################################################################
##                                                                                                                                     ##
##                                                    ANSWER TO QUESTION 4                                                             ##
##                                                                                                                                     ##
#########################################################################################################################################
#########################################################################################################################################

# install.packages("corrplot")
# install.packages("sjPlot")
pairs(usdata[,1:4])
par(mfrow = c(1,1))
library(corrplot)
corrplot(cor(usdata[,1:4]), method = 'ellipse')
library(sjPlot)
sjp.corr(usdata[,1:4], corr.method = "pearson", sort.corr = T)
# Correlation of price and feats (number of 11 features a house has)
plot(usdata[,4], usdata[,1], xlab=names(usdata)[4], ylab='Price',cex.lab=1.5)
abline(lm(usdata[,1]~usdata[,4]),col=2)
boxplot(usdata[,1]~usdata[,4], xlab=names(usdata)[4], ylab='Price',cex.lab=1.5)
abline(lm(usdata[,1]~usdata[,4]),col=2)
# Correlation of price (our response variable) and factor variables
par(mfrow=c(1,2))
for(j in 5:6){
  boxplot(usdata[,1]~usdata[,j], xlab=names(usdata)[j], ylab='Price',cex.lab=2.0)
}
# Correlations valid on the numerical variables
round(cor(usdata[,1:4]),2)
cor(usdata$PRICE, usdata$SQMT, method = "pearson")
# It seems to exist a linear relationship between price and sqmt (size of house in square meters)
```

```
##############################################################################################################################################
##############################################################################################################################################
##                                                                                                                                          ##
##                                                    ANSWER TO QUESTION 5                                                                  ##
##                                                                                                                                          ##
##############################################################################################################################################
##############################################################################################################################################

model <- lm(PRICE ~., data = usdata)
summary(model)
# R^2 is above 0.7, so current linear model is a good fit model, but not a very good (below 0.9)


##############################################################################################################################################
##############################################################################################################################################
##                                                                                                                                          ##
##                                                    ANSWER TO QUESTION 6                                                                  ##
##                                                                                                                                          ##
##############################################################################################################################################
##############################################################################################################################################

mfull <- lm(PRICE~.,data=usdata)
step(mfull, direction='both')
step(mfull, direction='both', k=log(100))
# SQMT and FEATS are the best variables for predicting PRICE
# PRICE = -17592.8 + 732.5*SQMT + 3983.7*FEATS


##############################################################################################################################################
##############################################################################################################################################
##                                                                                                                                          ##
##                                                    ANSWER TO QUESTION 7                                                                  ##
##                                                                                                                                          ##
##############################################################################################################################################
##############################################################################################################################################

round(sapply(usdata[,c(2,4)], mean),2)
round(sapply(usdata[,c(2,4)], sd),2)
model2 <- lm(PRICE ~ 1 + SQMT + FEATS, data = usdata)
# model2 <- lm(PRICE ~ . - AGE - NE - COR, data = usdata)
summary(model2)
# PRICE = -17592.8 + 732.5*SQMT + 3983.7*FEATS + ε, where ε ~ N(0,14370^2)
# model with centered covariates
usdata2 <- as.data.frame(scale(usdata[,1:4], center = TRUE, scale = F))
usdata2$NE <- usdata$NE
usdata2$COR <- usdata$COR
usdata2$PRICE <- usdata$PRICE
round(sapply(usdata2[,c(2,4)], mean),2)
round(sapply(usdata2[,c(2,4)], sd),2)
mfull2 <- lm(PRICE~.,data=usdata2)
step(mfull2, direction='both')
model3 <- lm(PRICE ~ 1 + SQMT + FEATS, data = usdata2)
summary(model3)
# PRICE = 115841.3 + 732.5*SQMT + 3983.7*FEATS + ε, where ε ~ N(0,14370^2)
anova(model2, model3)

##############################################################################################################################################
##############################################################################################################################################
##                                                                                                                                          ##
##                                                    ANSWER TO QUESTION 8                                                                  ##
##                                                                                                                                          ##
##############################################################################################################################################
##############################################################################################################################################

# install.packages("car")
# install.packages("randtests")
# install.packages("lmtest")
par(mfrow=c(1,1))
# Normality of errors (Q-Q Plot)
plot(model3, which = 2)
# Constant Variance
plot(model3, which = 3)
Stud.residuals <- rstudent(model3)
yhat <- fitted(model3)
par(mfrow=c(1,2))
plot(yhat, Stud.residuals)
abline(h=c(-2,2), col=2, lty=2)
plot(yhat, Stud.residuals^2)
abline(h=4, col=2, lty=2)
library(car)
ncvTest(model3)
yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab=6)
table(yhat.quantiles)
leveneTest(rstudent(model3)~yhat.quantiles)
par(mfrow=c(1,1))
boxplot(rstudent(model3)~yhat.quantiles)
# Non Linearity
residualPlot(model3, type='rstudent')
residualPlots(model3, plot=F, type = "rstudent")
# Independence of Errors
plot(rstudent(model3), type='l')
library(randtests)
randtests::runs.test(model3$res)
library(lmtest)
dwtest(model3)
library(car)
```

```
durbinWatsonTest(model3)


##############################################################################################################################################
##############################################################################################################################################
##                                                                                                                                          ##
##                                                        ANSWER TO QUESTION 9                                                              ##
##                                                                                                                                          ##
##############################################################################################################################################
##############################################################################################################################################

# install.packages("glmnet")
library(glmnet)
X <- model.matrix(mfull2)[,-1]
lasso <- glmnet(X, usdata2$PRICE)
plot(lasso, xvar = "lambda", label = T)
#Use cross validation to find a reasonable value for lambda
lasso1 <- cv.glmnet(X, usdata2$PRICE, alpha = 1)
lasso1$lambda
lasso1$lambda.min
lasso1$lambda.1se
plot(lasso1)
coef(lasso1, s = "lambda.min")
coef(lasso1, s = "lambda.1se")
plot(lasso1$glmnet.fit, xvar = "lambda", label = T)
abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)
```